# Biases within databases

Specifically sentiment analysis within two datasets

# Why?

Sentiment analysis provides a good starting point for making programs that can research.

Biases introduce sources of error.

| | date | polarity | query | text | user |
|---|---|---|---|---|---|
| 0 | Mon Jun 01 18:08:26 PDT 2009 | 4 | NO_QUERY | i'm 10x cooler than all of you! | katie4593 |
| 1 | Mon Jun 01 23:55:43 PDT 2009 | 0 | NO_QUERY | O.kk? Thats weird I cant stop following people on twitter... I have tons of people to unfollow | migaruler |
| 2 | Mon May 04 06:08:51 PDT 2009 | 4 | NO_QUERY | what a beautiful day not to got to my first class | ocean_waves301 |
| 3 | Sun May 31 18:42:57 PDT 2009 | 4 | NO_QUERY | .@HildyGottlieb & I was just saying to Maha'al yesterday, everything we ever needed to know was in Beatles' lyrics - you prove my point! | TerraScene |
| 4 | Sat May 09 18:35:44 PDT 2009 | 0 | NO_QUERY | kinda sad and confused why do guys do this? | jenny0404 |

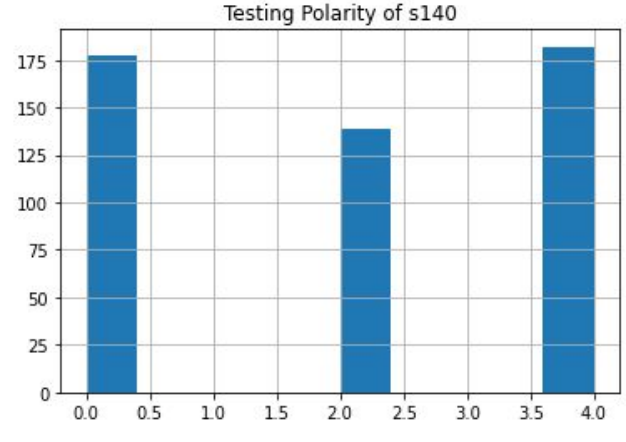
| | label | text |
|---|---|---|
| 0 | 0 (neg) | This was an absolutely terrible movie. Don't be lured in by Christopher Walken or Michael Ironside. Both are great actors, but this must simply be their worst role in history. Even their great acting could not redeem this movie's ridiculous storyline. This movie is an early nineties US propaganda piece. The most pathetic scenes were those when the Columbian rebels were making their cases for revolutions. Maria Conchita Alonso appeared phony, and her pseudo-love affair with Walken was nothing but a pathetic emotional plug in a movie that was devoid of any real meaning. I am disappointed that there are movies like this, ruining actor's like Christopher Walken's good name. I could barely sit through it. |
| 1 | 0 (neg) | I have been known to fall asleep during films, but this is usually due to a combination of things including, really tired, being warm and comfortable on the sette and having just eaten a lot. However on this occasion I fell asleep because the film was rubbish. The plot development was constant. Constantly slow and boring. Things seemed to happen, but with no explanation of what was causing them or why. I admit, I may have missed part of the film, but i watched the majority of it and everything just seemed to happen of its own accord without any real concern for anything else. I cant recommend this film at all. |
| 2 | 0 (neg) | Mann photographs the Alberta Rocky Mountains in a superb fashion, and Jimmy Stewart and Walter Brennan give enjoyable performances as they always seem to do. <br /><br />But come on Hollywood - a Mountie telling the people of Dawson City, Yukon to elect themselves a marshal (yes a marshal!) and to enforce the law themselves, then gunfighters battling it out on the streets for control of the town? <br /><br />Nothing even remotely resembling that happened on the Canadian side of the border during the Klondike gold rush. Mr. Mann and company appear to have mistaken Dawson City for Deadwood, the Canadian North for the American Wild West.<br /><br />Canadian viewers be prepared for a Reefer Madness type of enjoyable howl with this ludicrous plot, or, to shake your head in disgust. |

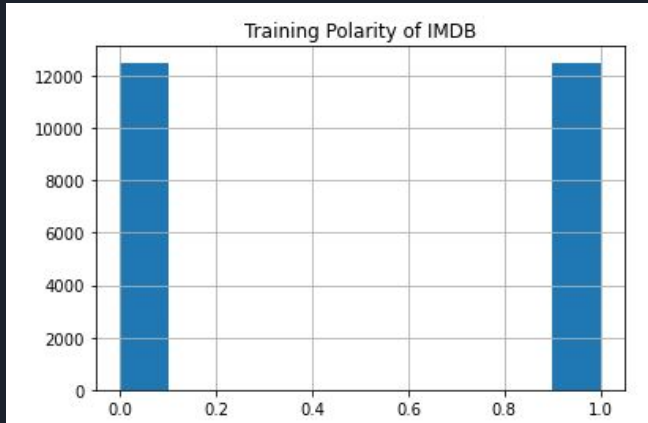
# Where I got the data

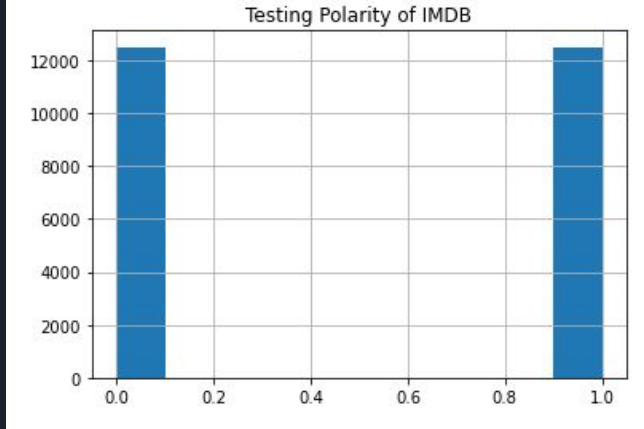
```
ds = tfds.load('sentiment140', as_supervised = True)
train_ds = ds['train']#these are used later for training and fitting
test_ds = ds['test']

imdb_ds = tfds.load('imdb_reviews', as_supervised = True)
imdb_train_ds = imdb_ds['train']
imdb_test_ds = imdb_ds['test']


ds = tfds.load('sentiment140')
train = ds['train']
test = ds['test']

ds = tfds.load('imdb_reviews')
imdb_train = ds['train']
imdb_test = ds['test']

train = tfds.as_dataframe(train)
test = tfds.as_dataframe(test)

imdb_train = tfds.as_dataframe(imdb_train)
imdb_test = tfds.as_dataframe(imdb_test)

train = train[['text', 'polarity']]
test = test[['text', 'polarity']]
```

```
t_ds = train_ds.map(lambda text, label: preprocess(text,label))
t_ds = train_ds.map(lambda text, label: (text, label/4))

imdb_train_ds = imdb_train_ds.map(lambda text,label: (text, -1 if label == 0 else label))
```

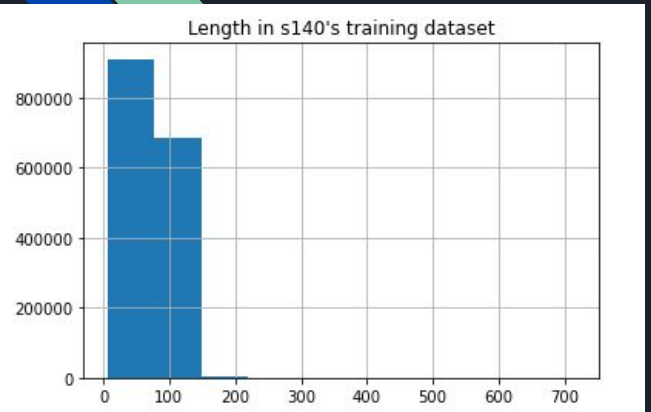
# Amount of each label

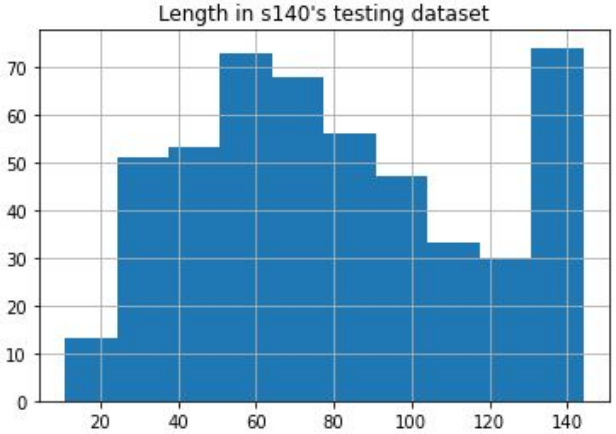



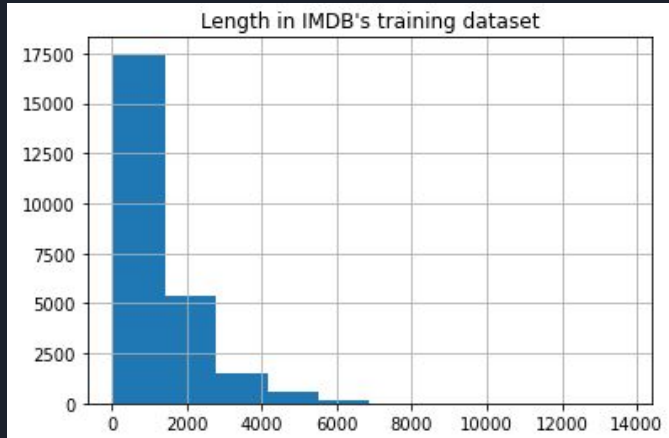



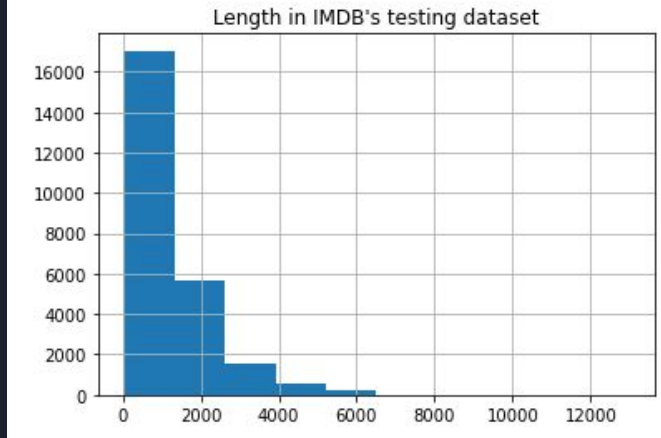



# The distribution of length

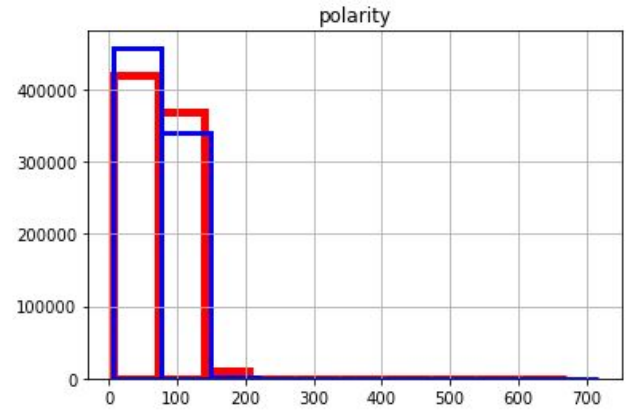


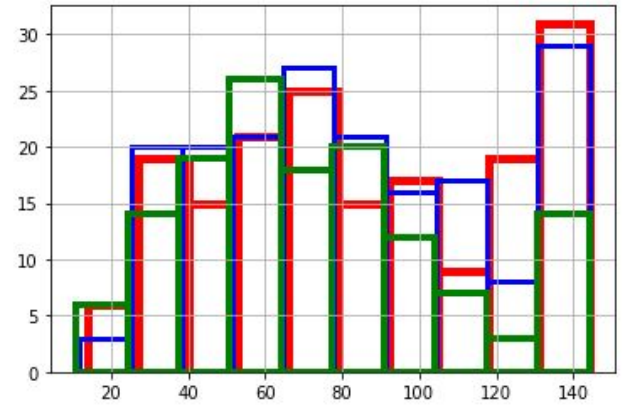



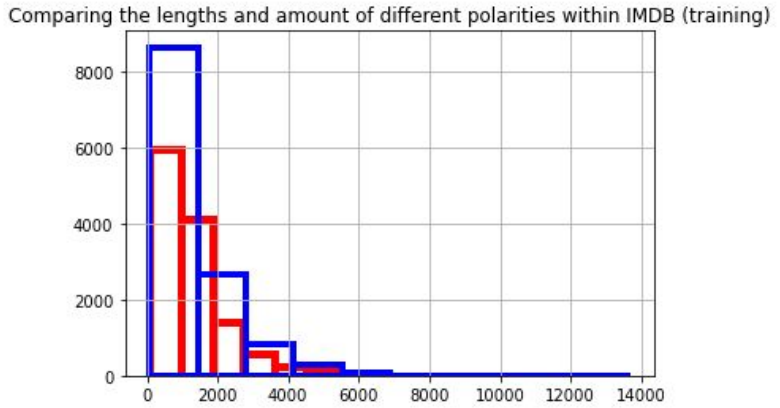



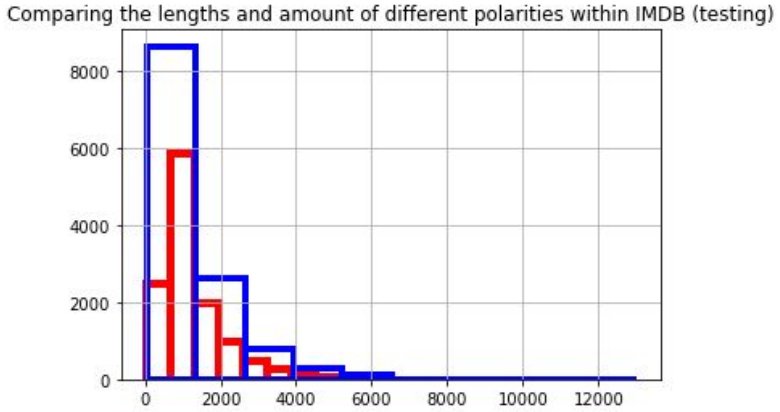
# Distribution of length based on labels



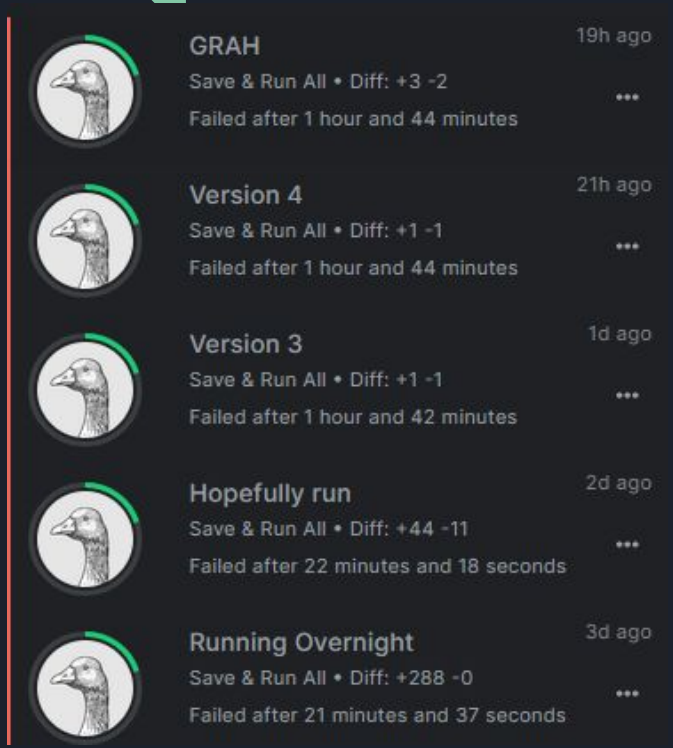




Red is negative ( 0 )
Green is neutral ( 2 )
Blue is positive ( 1 or 4 )



# Problems :(




```
h2 = imdb_model.fit(imdb_train.batch(128), epochs = 5, validation_data = test.batch(128))
```

```
---------------------------------------------------------------------------
AttributeError                            Traceback (most recent call last)
/tmp/ipykernel_23/2108579831.py in <module>
----> 1 h2 = imdb_model.fit(imdb_train.batch(128), epochs = 5, validation_data = test.batc
h(128))

/opt/conda/lib/python3.7/site-packages/pandas/core/generic.py in __getattr__(self, name)
   5485         ):
   5486             return self[name]
-> 5487         return object.__getattribute__(self, name)
   5488
   5489     def __setattr__(self, name: str, value) -> None:

AttributeError: 'StyledDataFrame' object has no attribute 'batch'
```

```
h2 = imdb_model.fit(imdb_train_ds.batch(128), epochs = 5, validation_data = test.batch(128))
```

```
---------------------------------------------------------------------------
AttributeError                            Traceback (most recent call last)
/tmp/ipykernel_23/3541868753.py in <module>
----> 1 h2 = imdb_model.fit(imdb_train_ds.batch(128), epochs = 5, validation_data = test.b
atch(128))

/opt/conda/lib/python3.7/site-packages/pandas/core/generic.py in __getattr__(self, name)
   5485         ):
   5486             return self[name]
-> 5487         return object.__getattribute__(self, name)
   5488
   5489     def __setattr__(self, name: str, value) -> None:

AttributeError: 'DataFrame' object has no attribute 'batch'
```

# Problems cont.

```
Minecraft is such a good game
the sentiment is [1.] for the sentiment140 model

I actually despise tall people
the sentiment is [1.] for the sentiment140 model

I love this show
the sentiment is [1.] for the sentiment140 model

I hate everyone and anyone who are more talented than me in any way.The Industrial Revolution and its consequences have
been a disaster for the human race. They have greatly
increased the life-expectancy of those of us who live in
“advanced” countries, but they have destabilized society,
have made life unfulfilling, have subjected human beings
to indignities, have led to widespread psychological suffe-
ring (in the Third World to physical suffering as well) and
have inflicted severe damage on the natural world. The
continued development of technology will worsen the si-
tuation. It will certainly subject human being to greater in-
dignities and inflict greater damage on the natural world,
it will probably lead to greater social disruption and psy-
chological suffering, and it may lead to increased physical
suffering even in “advanced” countries.
the sentiment is [1.] for the sentiment140 model
...
```

```
model.fit(s140.batch(8000), epochs=1, verbose = 1, workers = 4, use_multiprocessing= True)
✓ 29m 7.3s
```

# Results

```
Minecraft is such a good game
the sentiment is [0.6759199] for the sentiment140 model

I actually despise tall people
the sentiment is [0.5597734] for the sentiment140 model

I love this show
the sentiment is [0.6210745] for the sentiment140 model

I hate everyone and anyone who are more talented than me in any wa
been a disaster for the human race. They have greatly
increased the life-expectancy of those of us who live in
“advanced” countries, but they have destabilized society,
have made life unfulfilling, have subjected human beings
to indignities, have led to widespread psychological suffe-
ring (in the Third World to physical suffering as well) and
have inflicted severe damage on the natural world. The
continued development of technology will worsen the si-
tuation. It will certainly subject human being to greater in-
dignities and inflict greater damage on the natural world,
it will probably lead to greater social disruption and psy-
chological suffering, and it may lead to increased physical
suffering even in “advanced” countries.
the sentiment is [0.9978436] for the sentiment140 model

Hello everyone,
I want to take a moment to talk about the power of motivation. Mot

No matter what your goals or aspirations may be, it all starts wit

But staying motivated isn't always easy. There will be times when

It's important to keep in mind that motivation is not something th

One of the best ways to stay motivated is to keep your eye on the

Another key to staying motivated is to surround yourself with posi

So remember, motivation is the key to success. It's what separates

Thank you.
the sentiment is [0.19850135] for the sentiment140 model


    What is wrong with people? How can anyone think that wearing s

Let me break it down for you: socks and sandals serve completely d

And let's talk about the aesthetic of this abomination. Socks and

On top of all that, wearing socks with sandals is just plain confu

And don't even get me started on the message that wearing socks wi

So please, for the love of all that is good and holy, stop wearing
the sentiment is [0.99795437] for the sentiment140 model


the sentiment is [0.5057712] for the sentiment140 model
```